DATABASE SERVICES FOR PHYSICS AT CERN WITH ORACLE 10G RAC AND ASM ON LINUX

OVERVIEW

Physics Database Services and its DBA team at CERN provide Oracle-based services to the high energy physics community. With an installed base of about 25 TB-sized DB clusters (up to 8 nodes and 6 TB per database as of July 2008), the services are architected to accommodate growth and fulfill database demands for data analysis generated by CERN's new accelerator, the **Large Hadron Collider**.

The main challenges the services have to address are: high availability for mission critical applications, performance and scalability of several multi-terabyte applications, contain HW and DB administration complexity and overall costs. Moreover CERN database services are part of a distributed database network connecting several large scientific institutions (Tier 1 sites).

Oracle 10g RAC and ASM on Linux has been chosen as the main platform to deploy the database services. Oracle RAC provides scalability by deploying a **cluster** of nodes build with commodity HW are used to load balance the DB workload against shared database storage (shared-everything clustering technology). At the same time Oracle **RAC provides high availability,** because the failure of one single cluster node does not bring down the service. Oracle ASM works as a volume manager and specialized cluster filesystem, allowing the use of low-cost storage arrays to build **scalable and highly available storage** solution for Oracle 10g. Oracle clusters are composed of a relatively large number of cluster nodes and storage elements, allocated on commodity (**cost-effective**) hardware with homogeneous characteristics. This has the additional advantage of **simplifying administration**, hardware provisioning and service growth. The database architecture deployed for Physics DB services embodies the key ideas of grid computing, as implemented by Oracle10g.

As of July 2008, the services run on 450 cores (125 servers), with 900 GB of RAM, 500 TB of raw disk space and have more than 1000 deployed schemas.

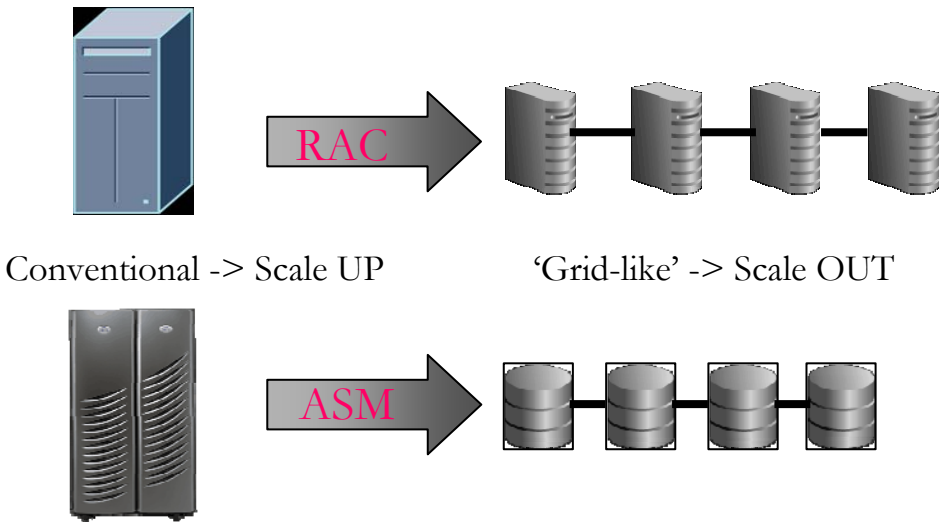

FIGURE 1: *Oracle 10g RAC and ASM allow to build clusters with a larger number of commodity HW components (scale-out) and achieve HA and performance goals previously available only with large SMP and enterprise storage solutions.*

ORACLE 10G RAC AND ASM FOR SCALABLE DATABASE SERVICES

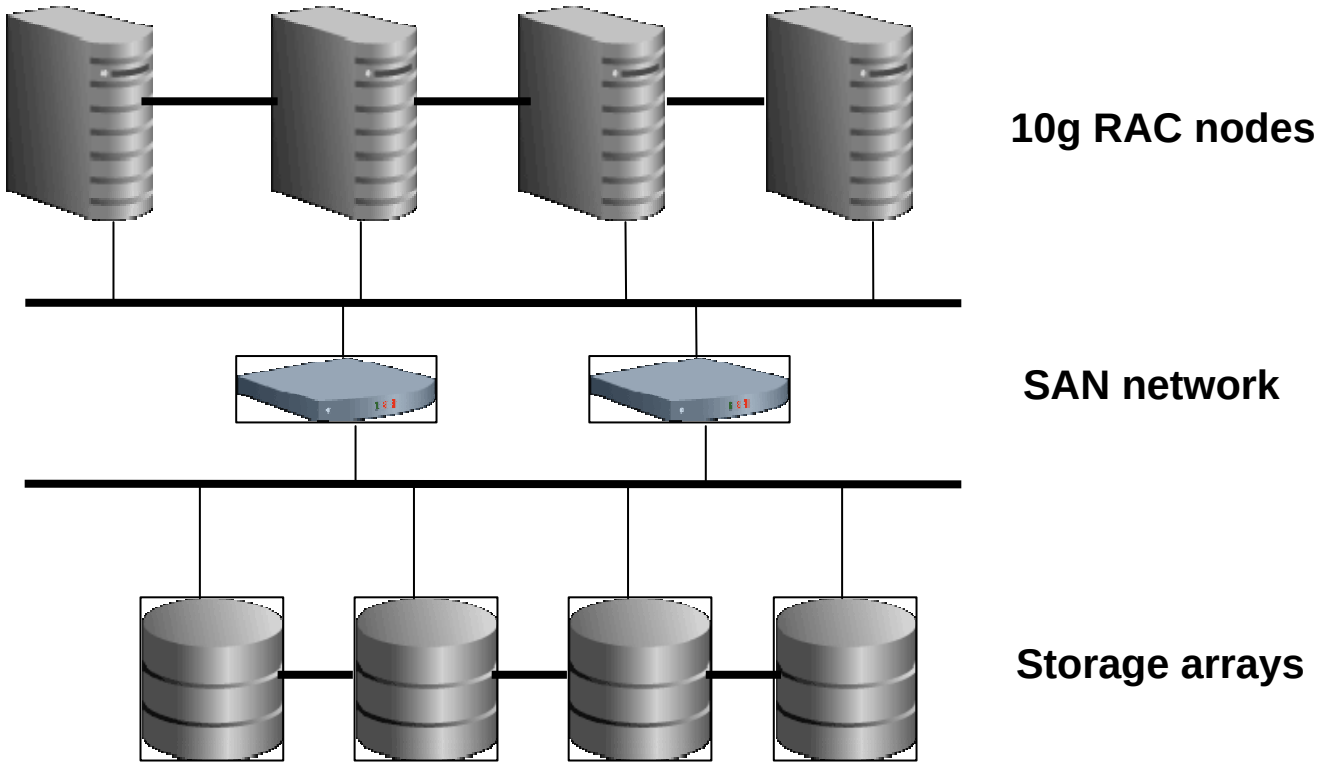

FIGURE 2: *Pictorial representation of a 4-node RAC with ASM on Oracle 10g for CERN's Physics DB services*

A scalable Oracle 10g architecture is currently deployed in production at CERN using the following key elements:

- Oracle 10g RAC on Linux to scale out Oracle workload (load balance cluster of servers).
- Fiber Channel SAN network
  - Storage can be reconfigured online using SAN zoning
  - SAN multipathing for resilience and load balancing
- Storage built with mid-range storage array with FC controllers and SATA HD
- Oracle ASM used as the volume manager and cluster filesystem for Oracle
  - storage striping and mirroring for performance and HA (similar to RAID 10)

FURTHER HW DETAILS AND SCALABILITY RESULTS

Oracle RAC production clusters are typically deployed over 4 nodes (and up to 8 nodes). Each node is a server with **two quad-core CPUs** (Intel E5410) running @ **2.33GHz** and with **16GB RAM**. Two **redundant cluster interconnects** are deployed with Gbps Ethernet. Public networks are also on Gbps Ethernet. The servers mount **HBAs that are dual-ported** and connected to redundant SAN switches. Access to the storage is configured via two (redundant) switch FC networks (4 Gbps networks are used on different clusters).

The disk arrays contain 'consumer-quality' SATA disks but have dual-ported Fiber Channel controllers. Disks are mapped from the storage array to the Linux servers directly as LUNs spanning whole physical disks (that is no RAID

configuration is used at the array level). Mirroring and striping are done instead at the software level with Oracle's ASM (**host based mirroring with ASM**). This has the advantage to allow mirroring across two different storage arrays (and therefore storage controllers). Storage arrays are dual-ported, where each port is connected to a different SAN switch.

The typical characteristics of the storage arrays used in production are: **16 SATA disks** with a raw capacity of 6 TB per array. Storage is configured such that each SATA disk in visible as a LUN to the Linux servers, then the disks are partitioned in 2 halves under Linux: the outer (faster) half of the disk is used to build data disk group, while the inner half is used to build the flash recovery area disk group. Disk groups are created with ASM and used to allocate Oracle files: data disk groups are used, among others, for datafiles, controlfiles and redologs. Flash recovery area disk groups are used to allocate the Oracle 10g flash recovery area, where, among others, archivelogs and backups to disks are stored.

The largest storage **configuration tested consisted in 26 storage arrays** of 16 disks each. A mirrored data diskgroup with a usable size of **70 TB was tested**. Furthermore the tests have shown **6 GByte/sec** sequential read performance and **40K small random reads IOPS.** The tested cluster was built with 14 servers (**112 cores**).

A study of the measured performances of the disk configuration as described here can be found in the PDB wiki pages (see links and references section further in this document).

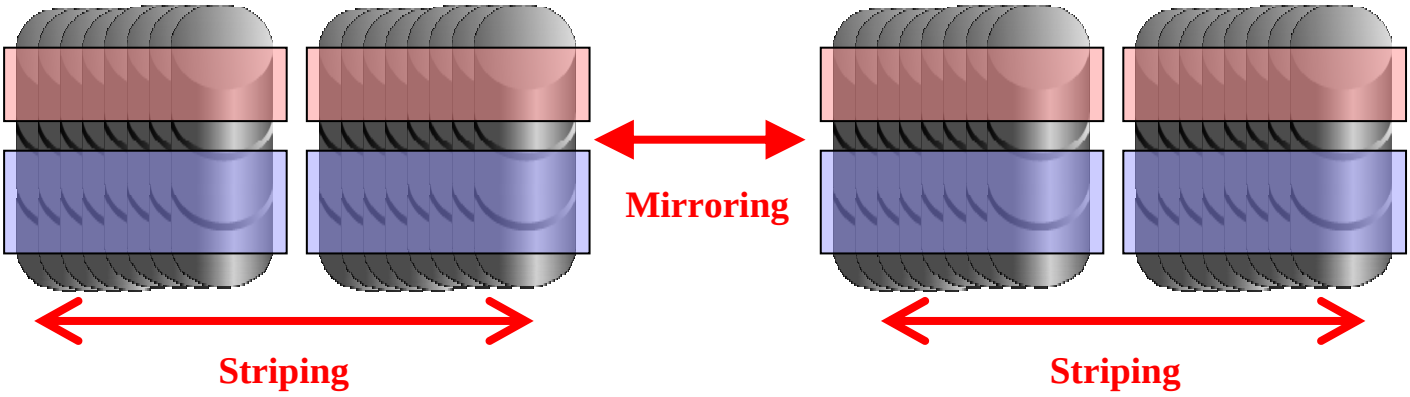

FIGURE 3: *ASM organizes storage into diskgroups, an implementation of 'SAME' (stripe and mirror everything) guidelines. Two diskgroups are shown in the figure (marked with pink and blue). Each diskgroup is built by striping and mirroring HD partitions. External disk partitions are used for the data diskgroup, internal partitions are used for the recovery disk group. This configuration is used when low-activity is measured on the recovery area, more active databases have complete separation of DATA nad RECORY areas.*

BACKUP STRATEGY

Tape backups are currently performed **both on tape and on disk**. Tape backups are performed with Oracle's RMAN and Tivoli storage manager. An incremental strategy is used because of the very large size of the databases: Level 0 (full) backups, level 1 cumulative (that is changes backup of all the changes since the latest level 0) and level 1 differential backups (that is all changes since the latest level 1 backup) are scheduled to optimize recovery time and at the same time reduce the load on the backup system. **A typical configuration is**: level 0 to tape every 2 weeks, level 1 cumulative every 3 days, level 1 differential daily.

Backus to disk (**flash backups**) have also been implemented to complement tape backups and to reduce recovery times for most failure scenarios. A copy of the database is kept in the flash recovery area and is refreshed daily. A copy of the archivelogs needed to recover the flash backup is also kept on disk. For example, in case multiple I/O failures or logical corruption of the production database, the DBA can perform a 'switch operation' to the flash copy on disk, then performs a recovery with the redologs and finally put the production back online. This operation can be performed in a relatively

small time window even for very large databases, while the full recovery of hardware and restore from tape backup can take several hours and/or days.

Flash backups come at basically no additional cost using the storage configuration described above. This is because of the large flash recovery areas that are allocated in the storage configuration described above. SATA disks used in the storage array are 400 GB in size each, for performance reasons (see links paragraph) only the outer 200 GB are used for data, while the rest is used for the flash recovery areas.

Flash backups are performed using **10g RMAN**. In particular Oracle 10g new features are leveraged to incrementally maintain flash backups for large databases (**incremental refresh of database copies** and **block change tracking** are two new features that make this possible). In this way the I/O resources needed are proportion to the amount of transactions and not to the overall size of the database.

SUMMARY

The Physics Database Services at CERN are deployed using Oracle 10g RAC and ASM on Linux. Clusters on Linux 'mid-range' servers are used to achieve the service goals in terms of high availability, scalability and consolidation. The hardware deployment model, where a large number of homogenous servers are bound together via Ethernet and FC networks, allows also for additional savings in provisioning, simplified management and a flexible architecture for growth.

LINKS AND REFERENCES

http://cern.ch/phydb - Physics Database Services Home Page

http://twiki.cern.ch/twiki/bin/view/PSSGroup/HAandPerf - Twiki page with this document and other links on performance and HA

http://twiki.cern.ch/twiki/pub/PSSGroup/Presentations2008/Storage_Studies_WLCG_workshop_25-4_LC.ppt - storage studies with 10g ASM and low-cost storage arrays

http://twiki.cern.ch/twiki/pub/PSSGroup/Presentations2006/UKOUG_RACSig_Oct06_LC.pdf - UKOUG RAC SIG presentation on CERN's oracle architecture

VERSIONS:

Feedback to: Luca.Canali@cern.ch

Major update, July 2008, L.C.

First update, Nov 2006, L.C.

First version - Jun 2006, L.C.